2019

Explanation < Justification: GDPR and the Perils of Privacy

Talia B. Gillis
Columbia Law School, gillis@law.columbia.edu

Josh Simons
Harvard University

ARTICLE

EXPLANATION < JUSTIFICATION: GDPR AND THE PERILS OF PRIVACY

TALIA B. GILLIS AND JOSH SIMONS[†]

The European Union's General Data Protection Regulation (GDPR) is the most comprehensive legislation yet enacted to govern algorithmic decision-making. Its reception has been dominated by a debate about whether it contains an individual right to an explanation of algorithmic decision-making. We argue that this debate is misguided in both the concepts it invokes and in its broader vision of accountability in modern democracies. It is justification that should guide approaches to governing algorithmic decision-making, not simply explanation. The form of justification – who is justifying what to whom – should determine the appropriate form of explanation. This suggests a sharper focus on systemic accountability, rather than technical explanations of models to isolated, rights-bearing individuals. We argue that the debate about the governance of algorithmic decision-making is hampered by its excessive focus on privacy. Moving beyond the privacy frame allows us to focus on institutions rather than individuals and on decision-making systems rather than the inner workings of algorithms. Future regulatory provisions should develop mechanisms within modern democracies to secure systemic accountability over time in the governance of algorithmic decision-making systems.

[†] Talia Gillis is an Empirical Law and Finance Fellow and SJD candidate at Harvard Law School, and a PhD candidate in Business Economics. Josh Simons is a PhD candidate in Government at Harvard University. He is a Graduate Fellow at the Edmond J. Safra Centre for Ethics and an Affiliate at the Berkman Klein Centre for Internet and Society.

INTRODUCTION

The European Union's General Data Protection Regulation (GDPR) is the most comprehensive legislation yet enacted to govern algorithmic decision-making. Its scope is supra-national, shaping the data protection practices of companies operating throughout the world's most prosperous integrated economic area. It establishes enforcement mechanisms with bite, threatening companies with fines of up to 4 percent of global turnover for the most serious violations. Yet the GDPR's focus is not decision-making, but privacy. This is the product of history. The primary protagonists of current debates about governing algorithmic decision-making are privacy scholars. We believe this privacy lens has distorted interpretations of the GDPR's approach to governing algorithmic decision-making. That approach reaches beyond an individual right to explanation, to establish provisions that aim to build systemic accountability over time.

This paper examines those provisions. We explore the tools the GDPR provides for ensuring that institutions justify their use of algorithmic decision-making systems, to both regulators and individuals subject to their decisions. Our aim is not simply to interpret the GDPR, though we side with scholars who argue that the main text of the GDPR must be read in conjunction with surrounding 'soft-law', including the Recitals, Article 29 Working Party (A29WP) guidance, and the interpretations of authorities mandated with enforcing its provisions.[1] Rather, our aim is to step back and examine the concepts that underpin the right to explanation debate, and the broader challenge of regulating algorithmic decision-making. We make three arguments.

[1] *See* Margot E. Kaminski, *Binary Governance: Lessons From the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. (forthcoming 2019), at 48; Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189, 197-199 (2019); Bryan Casey et al., *Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L.J. 143 (2019).

First, we argue that accountability is the foundational goal that should guide approaches to governing algorithmic decision-making. Accountability is achieved when an institution must justify its choices about how it developed and implemented its decision-making procedure, including the use of statistical techniques or machine learning, to an individual or institution with meaningful powers of oversight and enforcement. Accountability produces instrumental benefits, including encouraging the use of decision-making procedures that are consistent and verifiable, and providing mechanisms for identifying and addressing discrimination and injustice.[2] However, we argue that accountability is the foundational goal because of its intrinsic, rather than its instrumental value. Accountability is constitutive of democratic self-governance. It is part of what it means for a citizenry to authorize in an ongoing way the complex decision-making systems whose recommendations shape their lives. Other goals discussed in the literature are all in some way means to securing accountability. Individual explanations of algorithmic systems are valuable if and when they enable institutions to justify those systems to individuals and regulators, but they may not always further this end.[3] Transparency may be necessary for some forms of accountability, but neither constitutes nor is entirely sufficient for accountability.[4] In other words, accountability requires justification and justification requires explanation. The form of each should determine the form of the others.

Second, we distinguish between different forms of justification required to attain systemic accountability and consider the appropriate form of explanation in each. Recent scholarship has debated whether a right to an explanation exists in the GDPR, and if so, what its content might be. We argue that the form this explanation should take must depend on the form of accountability being pursued. By separating out different forms of justification, we set out how a 'right to explanation' (a "RtE") might further the aim of accountability, and how it might

---

[2] Jeremy Waldron, *Accountability: Fundamental to Democracy* 26 (NYU Pub. Law & Legal Theory Research Paper Series, Working Paper No. 14-13, Apr. 2014), https://papers.ssrn.com/abstract=2410812; MATTHEW V. FLINDERS, THE POLITICS OF ACCOUNTABILITY IN THE MODERN STATE (Aldershot: Ashgate 2001); ADAM PRZEWORSKI, SUSAN CAROL STOKES, AND BERNARD MANIN, DEMOCRACY, ACCOUNTABILITY, AND REPRESENTATION (Cambridge Univ. Press 1999).

[3] *See, e.g.,* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018).

[4] *See generally* Mike Ananny & Kate Crawford, *Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973 (2016); Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503, 1530 (2013).

hinder it. We argue that the technical explanation of a statistical or machine learning model is not sufficient for an institution to justify its decision-making procedure. Furthermore, we argue that such a technical explanation may even distract from the most important provisions of the GDPR for securing systemic accountability.[5] These include two crucial components. First, a range of mechanisms for ensuring that institutions justify their choices in the design and implementation of algorithmic decision-making systems – the critical *ex ante* stage in machine learning – including their broader policy and commercial aims. Second, that these mechanisms ensure justifications are offered to regulators with the necessary information, resources and powers, not simply isolated, rights-bearing individuals with limited information and expertise.

Third, we argue that the GDPR's focus on privacy underpins some of its most significant limits. We identify three such limits, some of which are about the law itself, others about recent interpretations of the law. First, recent interpretations of the law mistakenly focus on the actual algorithms and machine learning models, rather than the broader policy and commercial environment in which they are deployed. The aims an institution has in designing and implementing an algorithmic decision-making system shape the workings of the algorithm or model itself, but receive far less attention, at least in the interpretive literature. Second, the law itself is constrained by its focus on individual rights. Machine learning, the most common form of algorithmic decision-making, makes information about the design and implementation about the overall system critical to exercising meaningful oversight. Information about individual decisions will not enable individuals to grasp of the nature of the system whose decisions shape their lives, or enhance their capacity to demand a justification from the powerful institutions that designed it. For related reasons, the notice and consent framework is not an adequate mechanism by which to ensure meaningful institutional accountability. Third, the GDPR and the literature surrounding it have no satisfactory account of how its provisions are to be subject to democratic oversight. Accountability matters because it is constitutive of democratic self-government. Future regulatory provisions must focus more directly on developing mechanisms within modern democracies that can secure accountability in the governance of algorithmic decision-making systems.

---

[5] *See, e.g.,* Lillian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 19, 65-67 (2017); Lilian Edwards & Michael Veale, *Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?*, 16 IEEE SEC. & PRIV. 46, 50 (2018).

The paper proceeds in two sections. The first contains our conceptual argument. We begin by arguing that accountability is the foundational goal that should guide approaches to governing algorithmic decision-making. Explanations are instrumentally valuable insofar as they enable the process of giving and receiving justifications that constitutes accountability in a democracy. The second draws out the implications of this argument for interpreting the GDPR and for approaches to governing algorithmic decision-making more broadly. We focus specifically on machine learning in this paper. Though we are interested more broadly in governance approaches to algorithmic decision-making, focusing specifically on machine learning draws attention to the most acute practical and theoretical challenges. We focus our discussion on governing the use of machine learning in the private rather than the public sector.

We end by setting out some of the ways in which the limits to recent interpretations of the GDPR are related to their framing in terms of privacy. The challenges we face when developing governance systems for algorithmic decision-making go beyond concerns that can usefully be understood in terms of privacy.

I. EXPLANATION → JUSTIFICATION → ACCOUNTABILITY

This section outlines our conceptual argument. First, we argue that accountability is the foundational goal. It should guide our interpretations of the GDPR. It should also drive the questions we pose, and the answers we advance, as we confront the broader task of developing a comprehensive approach to governing algorithmic decision-making. Second, we consider the implications this has for the other concepts invoked in the debate about whether a right to explanation exists in the GDPR (hereafter the RtE debate). The most obvious is explanation itself, providing explanations of the logic of a machine learning model to ensure that its operation is, in some way, comprehensible to external human observers. Explanations are said to be valuable because there is something inherently important about individuals understanding the systems to which they are subject, that is, because they respect individual autonomy; and also because such understanding is instrumentally important, for individuals to challenge decisions or to identify bias and discrimination.[6] We argue that

[6] *See* Gianclaudio Malgieri & Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 INT'L DATA PRIV. L. 243, 250 (2017); Selbst & Barocas, *supra* note 3, at 40-46.

explanations of machine learning models are valuable if and when they are a means to provide justifications of the broader decision-making procedure. What matters is justifying why the rules are the way they are; explaining what the rules are must further this end.

The focus on individual, technical explanations has been driven by an uncritical bent towards transparency. Transparency is thought to matter because to see is to know, and knowledge is power. Transparency provides the information required for governance and oversight.[7] This is a mistake. Like explanation, transparency is an instrumental good. Transparency matters if and when it is required to further the aim of systemic accountability. These concepts are important not only for the RtE debate, but for thinking more broadly about the central aims that should guide any approach to governing algorithmic decision-making. This section aims to make progress towards such conceptual clarity.

A. *Accountability*

Accountability, we submit, is the foundational concept. It is the motivation that drives arguments for transparency and for various forms of explanation in machine learning. It should be the central aim of all approaches to governing decision-making using machine learning. It is therefore important to be clear about what accountability is and why it is valuable.

Accountability is about vertical power. Accountability empowers those who might otherwise be powerless, demanding that those who wield power over them offer an account of their conduct. In the modern world, its most familiar form is democratic accountability, in which those who control the apparatus of well-organized territorial states must offer an account of their conduct to citizens subject to their power. Democratic accountability, as Jeremy Waldron puts it, confers "authority on those who are otherwise powerless over those who are well endowed with power."[8] More generally, accountability can be said to pertain in the following structure. Party A is accountable to party B with respect to its conduct C, if A has an obligation to provide B with some justification for C, and may face some form of sanction if B finds A's justification to be inadequate.[9]

---

[7] *See* Ananny & Crawford, *supra* note 4, at 974-977; Zarsky, *supra* note 4, at 1533; *See generally,* DAVID BRIN, THE TRANSPARENT SOCIETY: WILL TECHNOLOGY FORCE US TO CHOOSE BETWEEN PRIVACY AND FREEDOM? (1998).

[8] *See* Waldron, *supra* note 2.

[9] *See* Reuben Binns, *Algorithmic Accountability and Public Reason*, 31 PHIL. & TECH. 543, 544 (2018); *see generally* MARK BOVENS ET AL., THE OXFORD HANDBOOK OF PUBLIC ACCOUNTABILITY (2014).

This is the principal-agent of accountability.[10] Accountability requires an agent, such as rulers, to justify their conduct to a principal, such as an electorate, subject to sanction through a range of mechanisms, most obviously, elections. The agent's exercise of power is shaped by the knowledge of the principal's inevitable judgement. Accountability ensures that those with power must justify their decisions to those who they will affect. Much like the threat of punishment, the idea is that this will change the behaviour of those decision-makers for the better.[11] To apply this view of accountability to decision-making procedures that use machine learning, let us suppose accountability pertains when: An institution (Party A) must justify its choices about how it developed and implemented its decision-making procedure (Conduct C), including the use of statistical techniques or machine learning, to an individual or institution with meaningful powers of oversight and enforcement (Party B).

Accountability can secure a range of instrumental benefits. It encourages institutions to use decision-making procedures that are consistent and verifiable, as consistency and verifiability tend to make for more persuasive justifications. It encourages institutions to identify discrimination in their decision-making procedures, and where possible, to address it in the design stage.[12] Structures of accountability can incentivize institutions to develop decision-making procedures with more care, consider a broad range of interests and perspectives, and evaluate more kinds of risk and possible harms.[13]

But accountability is about more than power. Part of the value of accountability is that it changes the conduct of those with power because they know that conduct will have to be justified. However, the more fundamental value of accountability is intrinsic. It is constitutive of democratic self-governance.[14] A king might fear the judgement of his

---

[10] This has been the dominant view of accountability explored in political science and political economy for the past two decades. *See generally* PRZEWORSKI ET AL., *supra* note 2; James D. Fearon, *Self-Enforcing Democracy*, 126 Q.J. ECON. 1661 (2011); FLINDERS, *supra* note 2; KAARE STRØM, WOLFGANG C. MÜLLER, AND TORBJÖRN BERGMAN, DELEGATION AND ACCOUNTABILITY IN PARLIAMENTARY DEMOCRACIES (2003).

[11] ROBERT D. BEHN, RETHINKING DEMOCRATIC ACCOUNTABILITY 3 (2001).

[12] Selbst & Barocas, *supra* note 3, at 42, 55.

[13] *See* Binns, *supra* note 9, at 547; Zarsky, *supra* note 4, at 1530-1550.

[14] This is part of Jeremy Waldron's argument, drawing on several recent critiques of the narrowness of the principal-agent approach to accountability, and considering its relationship to democracy more broadly. Waldron, *supra* note 2. *See also* JOHAN P. OLSEN, DEMOCRATIC ACCOUNTABILITY, POLITICAL ORDER, AND CHANGE: EXPLORING ACCOUNTABILITY PROCESSES IN AN ERA OF EUROPEAN TRANSFORMATION (2017); CRAIG T. BOROWIAK, ACCOUNTABILITY & DEMOCRACY: THE PITFALLS AND PROMISE OF POPULAR CONTROL (2011); Alexander H. Trechsel, *Reflexive Accountability and Direct Democracy*, 33 W. EUR. POL. 1050 (2010).

subjects. He might fear rebellion or resistance. The anticipation of that rebellion or resistance might shape the decisions he makes. But this is not accountability. The King need not justify his decisions; he has no obligation to offer an account of the decisions he has made, or his reasons for making them, to his subjects. Whereas in a democracy, as Waldron argues, "the accountable agents of the people owe the people an account of what they have been doing, and a refusal to provide this is simple insolence."[15]

Accountability is part of the practice of modern democracy. The giving and receiving of justifications is part of what it means to jointly govern ourselves. The agents who give and receive justifications are varied: sometimes individual citizens justify what they do or decide to other individual citizens, sometimes institutions justify what they do or decide to individual citizens, sometimes institutions justify what they do or decide to other institutions.[16] The content of their justifications might be varied too, including important decision-making processes and procedures that shape the lives of citizens. This broader view of accountability extends beyond the public realm. The most obvious form of accountability in a democracy is certainly the justification by public bodies of their conduct to citizens. But the rules and procedures that shape our collective future go far beyond those authored in the public realm. We expect companies who deliver important services to justify their decisions and procedures, to us as citizens, and to governments as our representatives. Facebook must justify how it moderates content to Congress.[17] Its content moderation system profoundly shapes how we interact as citizens; decisions about how that system works are of public concern; therefore, Facebook must justify those decisions to us, the public, or to our representatives.

Democracy and accountability are not, however, the same thing.[18] There may actually be important tensions between democracy and accountability. Mechanisms for accountability are often solutions to the problem of control – they need not, and often are not, democratic. Central banks and financial regulators are institutions of accountability, that is,

---

[15] Waldron, *supra* note 2, at 28.

[16] We side with Waldron on this point: whether the justification is offered or received by an individual, a multitude, or a legal corporation doesn't matter as much as some suppose. *Id.*

[17] *See* Kate Klonick, *Facebook Released Its Content Moderation Rules. Now What?*, N.Y. TIMES (April 26, 2018), https://www.nytimes.com/2018/04/26/opinion/facebook-content-moderation-rules.html.

[18] Mechanisms of accountability may actually change how we do democracy. If accountability changes democracy, the two cannot be synonymous. *See generally* OLSEN, *supra* note 14; Trechsel, *supra* note 14.

they solve the problem of control, but they are not democratic. Environmental regulators are institutions of accountability, but they are not democratic. As Borowiak puts it, "accountability institutions can create veils of legitimacy that mask abuses and dampen the critical and participatory energies of the public. So doing, they can end up thwarting citizen control rather than enhancing it."[19] These are important issues, but we want to put them aside. The challenge of ensuring that institutions of accountability do not erode the possibilities of democratic action and legitimacy is critical to the future of democracy in increasingly complex societies and economies, but it is a separate challenge to thinking systematically about accountability and what is required to achieve it.

Accountability, then, is the foundational concept. What follows, we believe, is that transparency must be put in its proper place. Transparency is valuable insofar as it furthers the aim of accountability.[20] The conditions in which transparency furthers this aim are more limited than it is often supposed. Demands for transparency tend to assume that if people are provided with the necessary information, they will take action against decisions they think are wrong. The GDPR, for example, requires individuals to be provided with "meaningful information about the logic involved" in the automated decision[21] as part of the right to contest these decisions[22] and to enforce other rights under the GDPR.[23]

There are good reasons to be deeply skeptical about the connection between the provision of information to individuals and those individuals taking desired actions. Firstly, people have to understand the information they receive. There is ample evidence that people struggle with even simple and straightforward disclosures,[24] let alone disclosure that pertains to more technical aspects of automated decision-making. Second, people

---

[19] *See* BOROWIAK, *supra* note 14, at 179.

[20] *See, e.g.,* Ananny and Crawford, *supra* note 4; Adrian Weller, *Challenges for Transparency*, CORNELL UNIV. (July 29, 2017), http://arxiv.org/abs/1708.01870; Danielle Citron, *What to Do about the Emerging Threat of Censorship Creep on the Internet,* CATO INST. (November 28, 2017), https://www.cato.org/publications/policy-analysis/what-do-about-emerging-threat-censorship-creep-internet.

[21] Namely Article 13(2)(f), Article 14(2)(g) and Article 15(1)(g). Regulation (EU) 2016/679 of the European Parliament and the Council of April 27, 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 [hereinafter GDPR].

[22] GDPR art. 22.

[23] *See* discussion in Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT. DATA PRIV. L. 233 (2017) (showing the connection between providing information and individuals enforcing their rights).

[24] *See* discussion in Talia B. Gillis, *Putting Disclosure to the Test: Toward Better Evidence Based Policy*, 28 LOY. CONSUMER L. REV. 31, 50 (2015).

must understand how that information relates to their particular circumstances and preferences. Many years of research on the effect of information disclosure, in multiple realms, demonstrate that there is a significant gap between the promise of disclosures and their actual impact.[25] The drive towards transparency often produces legal and policy regimes that fail to achieve genuine accountability over time.

Accountability should be the central aim of all approaches to governing decision-making using machine learning. The giving and receiving of justifications is part of what it means for a citizenry to authorise in an ongoing way the complex decision-making systems whose recommendations shape their lives.

B. *Explanation < Justification*

We now turn to the central concept in the RtE debate, on which most interpretations of the GDPR have focused: explanation. On the face of it, the role of explanation in our notion of accountability seems obvious. Accountability requires justification and justification requires explanation. To justify a decision-making procedure that involves or is constituted by a machine learning model, an individual subject to that decision-making procedure requires an explanation of how the machine learning model works. This is the thought that underpins much of the RtE debate.

But let's pause for a moment to ask: What form of explanation does justification require? Think of an example. Suppose you are involved in a major car crash that leaves you paralyzed from the waist down. After you wake up in hospital, you ask: Why did I crash? The crash investigator helpfully left a report by the side of your bed. It explains: The velocity of your car produced a centrifugal force on your wheel hub, which, gradually produced a rotating motion on your wheel stud which, in turn, loosened your front left wheel from your chassis. The resulting force made your vehicle swerve to the left. The particles of the central barrier then came into contact

---

[25] *See generally* ARCHON FUNG ET AL., FULL DISCLOSURE: THE PERILS AND PROMISE OF TRANSPARENCY (2007); Lauren Willis, *The Consumer Financial Protection Bureau and the Quest for Consumer Comprehension*, 3 RUSSELL SAGE FOUND. J. SOC. SCI. 74 (2017); OMRI BEN-SHAHAR & CARL E. SCHNEIDER, MORE THAN YOU WANTED TO KNOW: THE FAILURE OF MANDATED DISCLOSURE (2014); Omri Ben-Shahar & Carl E. Schneider, *The Failure of Mandated Disclosure*, 159 U. PA. L. REV. 647 (2011); Ryan Bubb, *TMI? Why the Optimal Architecture of Disclosure Remains TBD*, 113 MICH. L. REV. 1021 (2015); Matthew A. Edwards, *Empirical and Behavioral Critiques of Mandatory Disclosure: Socio-Economics and the Quest for Truth in Lending*, 14 CORNELL J. L. & PUB. POL'Y 199, 229 (2004). For further analysis of the ideology underlying calls for transparency, s*ee* David E. Pozen, *Transparency's Ideological Drift*, 128 YALE L.J. 100 (2018) (arguing that transparency has shifted from an idea that promotes a stronger and more egalitarian regulatory state, to a tool aimed at limiting government intervention and regulation).

with the polymers on the left side of your vehicle. The molecular structure of the polymer was broken on the driver's side, rapidly reducing the speed of your vehicle and eventually bringing it to a halt. This explanation is clearly unsatisfactory. It's an explanation at the wrong level. It answers your 'why' question with an account of microphysics. You want to know why your wheel came off. The explanation might be true, but it is beside the point. What you *really* want is for Ford to justify why your wheel came off despite having serviced your vehicle last month. What matters is the justification that is part of the process of accountability. The form of explanation involved in that justification depends on the context. If Ford sent you an account of the microphysics of your crash, you would consider that not just a misunderstanding about the information you require, but an evasion of accountability. It represents a failure to justify what happened.

The RtE debate often conceives of explanations at completely the wrong level. More precisely, at a level that is simply not relevant to justification, and therefore, to accountability. To those subject to the decisions of a machine learning model, offering an explanation of a machine learning model is a little like offering an account of microphysics to explain a car crash. Explanations of machine learning models are certainly not sufficient for many of the most important forms of justification in modern democracies, and often, they may not even be necessary. More specifically, what form of explanation is necessary, including whether a technical explanation of the machine learning model is necessary, depends on who is justifying what to whom. This implies two important shifts in focus. First, in terms of *what* is being justified. The focus should be on how institutions justify their choices about the design and integration of machine learning models into their decision-making systems, rather than on what the rules governing a model's operation are. What matters is why the rules of an algorithm are what they are.[26] Second, in terms of *to whom* the justification is being offered. Institutions should justify their choices about the design and integration of machine learning models not to individuals, but to empowered regulators or other forms of public oversight bodies. Less emphasis should be placed on the rights of disempowered and isolated individuals, who are expected to understand the complicated models to whose decisions they are subject, and more on systemic accountability – the way power is structured between institutions. If accountability is the foundational goal, what is required is institutional justification, not algorithmic explanation. Algorithmic explanation can be necessary to institutional justification. But since it is justification that is necessary for accountability, and it is accountability that is of ultimate importance, the

[26] Selbst & Barocas, *supra* note 3.

form of institutional justification should determine the appropriate form of explanation.

The excessive focus on technical forms of explanation is itself the result of an uncritical bent towards transparency. This is partly the product of history. Much of the policy and legal debate about the governance of machine learning has developed from older debates about privacy. Many scholars who were previously privacy experts now write about the governance of artificial intelligence. The GDPR is framed as a privacy law, even though its focus reaches far beyond the confines of privacy.[27] The transparency bent, with all its pitfalls, has been unreflectively transposed from the privacy literature to the literature on explanation and interpretability.[28] The risk is that the limits of the transparency debate swiftly become limits to the debate about how we should integrate machine learning models into some of our most important social, economic and political institutions. The most important limit is the focus on individual rights, rather than on structures of power. The privacy debate has always been hemmed in by its focus on individual consent, a concept that has proved to be a mirage in theory and in practice.[29] As a result, it has overlooked more fundamental and intractable challenges about how institutions should hold one another to account, most notably, involving questions about the structure and distribution of power. If individual-understanding-of-machine-learning-models becomes the new individual-consent-to-the-use-of-their-data, we should expect a wholesale failure to hold to account the institutions that use machine learning for their own ends.

This uncritical bent towards transparency, and the subsequent focus on technical explanation, actually *suits* many of the most powerful actors in the internet age. The focus on algorithmic explanation can deflect from the need for institutional justification. Consider an example. To satisfy increasing calls for oversight and accountability in content moderation, suppose Facebook rolls out a new interactive tool. This tool allows individual users to interact with their News Feed, to understand the factors that 'explain' why they see particular pieces of content. Users would be able to change important parameters about themselves, such as their gender, race, or

---

[27] For instance, much of the literature about explanation in law is published in journals that are putatively about privacy. *See generally* Edwards & Veale, *Enslaving the Algorithm*, *supra* note 5.

[28] *See generally* BRIN, *supra* note 7; Will Thomas Devries, *Protecting Privacy in the Digital Age*, 18 BERKELEY TECH. L.J. 283, 283–311 (2003); Joshua A. T. Fairfield & Christoph Engel, *Privacy as a Public Good*, 65 DUKE L.J. 385, 385–457 (2016).

[29] Lothar Determann, *Social Media Privacy: A Dozen Myths and Facts*, 16 STAN. TECH. L. REV. 1, 7-10 (2012); Dan Svirsky, *Why Are Privacy Preferences Inconsistent?* 15 (JOHN M. OLIN CTR. FOR LAW, ECON., & BUS. FELLOWS' DISCUSSION PAPER SERIES, HARV. LAW SCH., Discussion Paper No. 81, 2018).

location, but also their behaviour on Facebook, such as what groups they have liked, or what publishers they read, and see how their News Feed changes. No doubt many users would feel Facebook had discharged its responsibility to explain how News Feed works. But this is not satisfactory. To the question "Why do I see what I see?", the tool effectively says "Well, if you were African American and not white, you'd see this; if you were female and not male, you'd see this; if you were from California and not Wisconsin, you'd see this; if you had a lower proportion of photos that contained cats, you'd see this", and so on. By implication, it says: "You see this because you are a white male from Wisconsin who likes cats." That explanation may be true. It may even enable a user to develop an intuitive picture of how Facebook's News Feed ranking systems work (though we are sceptical even of that, since that intuitive picture applies only to their case and may not generalize).[30] But it is nonetheless beside the point. The individual wants to know why Facebook chose to construct its News Feed ranking system in the way it did. Why are engagement and relevance the primary metrics, and how are they defined? What are the other principles on the basis of which content is promoted and demoted on News Feed? They want Facebook to justify its News Feed ranking system. The kind of explanation this requires is on the level of choices and principles in the design of content moderation systems, not of interpretable machine learning models. Such technical explanations can actually distract from the appropriate form of justification. Citizens feel they no longer need to press for answers to the harder, but more fundamental question: Why do you distribute information in this way?

The posing of these questions by citizens, and the answering of them by institutions, is essential to the functioning of modern democracies. For large internet companies in particular, the drive towards transparency, cashed out in the form of the search for interpretable machine learning models, represents a welcome distraction from a fundamental debate about their own powers and purposes. The danger is that we make the same mistake in explanation and interpretability as we have in privacy: individual 'understanding' of a model takes the role 'consent' is supposed to play in securing important forms of institutional accountability. Individual understanding may often be just as much of an illusion as individual consent.[31]

---

[30] Along the lines of the kind of interactive explanation systems about which Edwards and Veale are more optimistic. *See generally* Edwards & Veale, *Slave to the Algorithm*, *supra* note 5.

[31] *See generally* Lilian Edwards, *Privacy, Law, Code and Social Networking Sites*, RESEARCH HANDBOOK ON GOVERNANCE OF THE INTERNET (2013); Rikke Frank Joergensen, *The Unbearable Lightness of User Consent*, 3 INTERNET POL'Y REV. (2014);

Accountability is constitutive of democratic self-governance. It is part of what it means for citizens to authorise in an ongoing way the complex decision-making procedures to which they are subject. Accountability requires that an institution justify its choices for the design and implementation of its decision-making procedures, including the use of statistical techniques and machine learning, to an individual or institution with meaningful powers of oversight and enforcement. The right form of explanation can be crucial to the giving and receiving of that justification. The wrong form can unintentionally or intentionally undermine it. Technical explanations of machine learning models can further the aim of institutional justification, and therefore of accountability. But they can also undermine and distract from it. The form of explanation should depend on the form of accountability. Institutional context should drive the form of explanation offered. We cannot simply adopt technical solutions to explanation without thinking through what is required for genuine accountability over time. It is to this challenge, and to the interpretation of the GDPR with this aim in mind, that we now turn.

II. SYSTEMIC ACCOUNTABILITY, JUSTIFICATION, AND THE GDPR

We now turn to the GDPR and the RtE debate that has dominated its reception amongst scholars in the U.S. The GDPR, which came into effect in May 2018, lays down requirements with respect to the information individuals must receive about automated decision-making in their case.[32] Several recent proposals have followed suit, seeking to ensure that machine learning models, which might otherwise be uninterpretable, can be explained to those whose lives they will

Elizabeth Denham, *Consent Is Not the 'Silver Bullet' for GDPR Compliance*, INFO. COMM'R OFF. NEWS BLOG (August 16, 2017), https://ico.org.uk/about-the-ico/news-and-events/blog-consent-is-not-the-silver-bullet-for-gdpr-compliance/.

[32] *See generally* Kaminski, *The Right to Explanation, Explained*, *supra* note 1; Casey et al., *Rethinking Explainable Machines*, *supra* note 1; Isak Mendoza & Lee Bygrave, *The Right Not to Be Subject to Automated Decisions Based on Profiling*, EU INTERNET L.: REG. AND ENFORCEMENT (2017); Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision Making and a "Right to Explanation,"* 38 AI MAG. 50, 50–57 (2017); Malgieri & Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, *supra* note 6; Selbst & Powles, *Meaningful Information and the Right to Explanation*, *supra* note 23; Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT. DATA PRIV. L. 76, 76–99 (2017); Sandra Wachter et al., *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841 (2018).

inevitably shape.[33] Broadly, the GDPR requires individuals to be provided with "meaningful information about the logic involved" in the automated decision[34] as part of the right to contest these decisions and to enforce other rights under the GDPR.[35]

There has been fierce disagreement about the scope and content of this explainability requirement. The core of the right to explanation in the GDPR regime can be found in Article 22 and Articles 13, 14, and 15. Article 22 lays down the general assumption against "automated individual decision-making, including profiling" and articulates the three exceptions to that assumption, while Article 13, Article 14 and Article 15 discuss the various transparency rights that arise from the use of automated decision-making, including the right to explanation. Article 13 creates requirements at the time information is collected from an individual, Article 14 focuses on requirements at the time information is collected from a third party, and Article 15 creates ongoing requirements related to the holding of individuals' information. These Articles bear on cases of decisions "based solely on automated processing" which "produce[s] legal effects concerning him or her or similarly significantly affects him or her," as they require the individual to be informed of the existence of the automated decision-making and for the provision of "meaningful information about the logic involved" in the automated decision.[36]

Our aim is not to offer another interpretation of this requirement. We agree with scholars who have recently argued that the GDPR's main text must be read alongside surrounding 'soft-law.'[37] These include the preamble to the Directive, known as the Recitals. These Recitals are not strictly binding, but they indicate how the GDPR is likely to be enforced and how, therefore, companies are likely to shape their behaviour to comply with the GDPR.[38] They also include the guidance of the Article

---

[33] *See* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017); Edwards & Veale, *Enslaving the Algorithm*, *supra* note 5; Edwards & Veale, *Slave to the Algorithm*, *supra* note 5.

[34] *See* GDPR arts. 13(2)(f), 14(2)(g) and 15(1)(g).

[35] Selbst and Powles make explicit this connection between providing information and individuals enforcing their rights. Selbst & Powles, *Meaningful Information and the Right to Explanation*, *supra* note 23.

[36] An individual also has the right to contest these decisions under Article 22. GDPR, *supra* note 21.

[37] *See generally* Kaminski, *Binary Governance*, *supra* note 1; Kaminski, *The Right to Explanation, Explained*, *supra* note 1; Casey et al., *Rethinking Explainable Machines*, *supra* note 1.

[38] As Kaminski puts it, "[t]hese texts are not technically binding, but they provide clarity of what is to come." Kaminski, *supra* note 1, at 195; In contrast, Wachter et al., who argue that the Recitals are not binding in the case of establishing the right to explanation since they are only

29 Working Party (A29WP), an advisory board made up of data protection authority representatives of all EU Member States, the European Data Protection Supervisor, and the European Commission. The purpose of the A29WP and its successor, the European Data Protection Board, is to promote consistent application of the GDPR across Member States.[39] Furthermore, the GDPR is designed to be given force by national Data Protection Authorities (DPAs), like many other EU Directives. How those institutions interpret its provisions is, in the end, what matters. We therefore give particular weight to guidance subsequently issued by national DPAs, most notably, the Information Commissioner's Office (ICO) in the U.K.[40]

In our view, this accompanying guidance makes it clear that the GDPR does contain a right to explanation. But more importantly, that guidance should shape how we elaborate on the content and scope of that right to explanation. The guidance suggests that the GDPR has begun to develop a comprehensive set of provisions for attaining systemic accountability over time. What a right to an explanation means in the context of the GDPR should depend on how the GDPR aims to secure systemic accountability.

Our aim is to approach the challenge of explainability by keeping in mind what is of ultimate importance: holding those with power to account, by ensuring that institutions justify their design and use of machine learning models to regulatory bodies and to individuals subject to their predictions, classifications, and rankings. The appropriate form of explanation should depend on who is justifying what to whom, as part of the process of accountability. To draw out the implications of this argument for interpreting the GDPR, we propose a simple taxonomy of justifications. It is broken down by three questions: (1) *Who* is offering

---

meant to provide guidance in cases of ambiguity, which is not the case, they argue, with respect to Article 22. Moreover, they argue that the Recital could not be used to establish new legal rights and duties that do not clearly exist in the text of the Directive. *See* Wachter, et al., *supra* note 32, at 80.

[39] GDPR replaced the pre-existing EU Directive on privacy, the Data Protection Directive, which came into force in 1995, and was suspended when the GDPR became enforceable in 2018.

[40] *See generally* ICO, *Guide to the General Data Protection Regulation (GDPR)*, U.K. INFO. COMM'R OFF. (August 2018), https://ico.org.uk/media/for-organisations/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf; *See also* ICO, *Rights Related to Automated Decision Making Including Profiling*, U.K. INFO. COMM'R OFF. (2017), https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling/; *see also* ICO, *Big Data, Artificial Intelligence, Machine Learning and Data Protection*, U.K. INFO. COMM'R OFF. (2017), https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

the justification?; (2) *What* is being justified?; and (3) *To whom* is the justification offered?

Who should offer the justification?

We leave this question to one side. It implies distinctions between both between engineers who design the model and the institutions that use it as part of their decision-making process, and between different forms of institutions, such as private and public. We focus on the justification of decision-making processes by private institutions.

What should be justified?

1. The machine learning model (overall logic or specific predictions); OR

2. The choices an institution makes about the design of a machine learning model and its integration into their decision-making procedure.

To whom should the justification be offered?

A. An individual subject to the model's predictions, classifications or rankings; OR

B. A regulator or some other type of public oversight body.

We focus on two categories of justification that can be drawn from this taxonomy. The first is 1A. The explanation of a machine learning model (1) to an individual (A). The second is 2B. The explanation of the choices an institution makes in the design and implementation of a decision-making procedure (2) to a regulator or some other public oversight body (B). Let's take each in turn.

The debate about whether the GDPR contains a RtE focuses on the 1A category. It concerns whether an individual has a right to "meaningful information about the logic involved" in a fully automated decision which "significantly affects him or her."[41] This has produced a range of approaches to explaining machine learning models to individuals that would satisfy this requirement, from straightforward

[41] Selbst & Powles, *Meaningful Information and the Right to Explanation, supra* note 23.

counter-factual explanations[42] to more complex technical approaches to developing interpretable models. These technical approaches aim to summarise the logic of a complex machine learning model in a simpler, more comprehensible model. Most explain how machine learning models work after the fact, known as reverse engineering. These tend to either summarise the whole logic of the model, known as global approaches, or to explain a specific set of outcomes the model produces, known as local approaches.[43]

We believe this focus on the IA category is mistaken. The IA category, the requirement that an institution explain how its machine learning model works to an individual subject to those decisions, is not a satisfactory way of holding institutions to account. Knowing what the rules are is not itself a check on the power of those who decide what the rules are. The category mistakenly characterises a challenge of institutional justification as a challenge of algorithmic explanation. Focusing on the requirement of those with power to inform subjects as to what the rules are, intentionally or not, distracts from the higher-order question of what the rules should be. If Facebook offers a tool that allows an individual to understand why their News Feed shows them what it does, the danger is that the user feels as though Facebook has justified its more general choices about how it distributes information on News Feed. It has in fact done nothing of the sort. It suits Facebook for the debate to focus on how they can develop technical explanations of News Feed's ranking models, rather than on the principles Facebook chooses to impose on its content moderation systems. The latter draws attention to Facebook's underlying power to decide who sees what, and why.

Nor is the IA category a satisfactory interpretation of the GDPR's most important provisions. The GDPR contains important mechanisms for systemic accountability, which focus on forcing an institution to

---

[42] *See* Wachter et al., *supra* note 32, at 854.

[43] *See generally* Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, 51 ACM COMPUTING SURVEYS 1, 1–42 (2018); Philip Adler et al., *Auditing Black-Box Models for Indirect Influence*, 54 KNOWLEDGE AND INFO. SYS. 95, 95–122 (2018); Selbst and Barocas *supra* note 3; Zachary Lipton, *The Mythos of Model Interpretability*, CORNELL UNIV. (2017), https://arxiv.org/abs/1606.03490; Kroll et al., *supra* note 33; Jatinder Singh et al., *Responsibility & Machine Learning: Part of a Process* (Working Paper, 2016), https://papers.ssrn.com/abstract=2860048; Marco T. Ribeiro et al., *Why Should I Trust You?: Explaining the Predictions of Any Classifier*, 22 ACM 1135, 1135–1144 (2016); Tameru Hailesilassie, *Rule Extraction Algorithm for Deep Neural Networks: A Review*, 14 INT'L J. COMP. SCI. & INFO. SEC. 376, 376–380 (2016); Anupam Datta et al. *Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems*, 37 IEEE SYMP. ON SEC. & PRIV. 598, 598–617 (2016).

justify their choices in the design and implementation of algorithmic decision-making systems, including their broader policy and commercial aims. Read in conjunction with the accompanying guidance of the Recitals, and the guidance published by A29WP and the ICO, the GDPR contains provisions that have the potential to transform the *ex-ante* process of designing machine learning models and integrating them into the decision-making systems of a range of important institutions. It sets out clear mechanisms for structuring systemic accountability, to ensure institutions justify the choices they make in that process. These include empowered DPAs, broad Data Protection Impact Assessments (DPIAs), auditing, and ethical review boards.[44]

This section uses our taxonomy of justifications to explore what this broader, more expansive reading of the GDPR implies for various forms of explanation. We contrast our 1A and 2B categories—the explanation of a machine learning model to an individual and the explanation of the decisions made in the design and implementation of that model to a regulator—to explore what is wrong with the more limited readings of the GDPR's provisions. The aim is to learn some broader lessons about the governance of institutions designing decision-making systems that use machine learning.

A. *What Should Be Justified: Institutions and the Process of Machine Learning*

We first focus on what it is that should be justified in the process of securing systemic accountability in the governance of algorithmic decision-making. If it is the machine learning model itself that must be justified, it would seem to follow that such a justification depends on an explanation of how the model works, either in terms of its overall logic or some subset of specific predictions.

This reasoning is mistaken, but it is encouraged by the text of the GDPR itself. Article 22 focuses on decisions "based solely" on automated data processing. The question of what exactly this means has divided scholars. Some have argued that decision-making procedures which involve humans in some perfunctory way would be exempt from Article

[44] *See* Kaminski, *The Right to Be Explained, Explained*, *supra* note 1, at 208 ("Accompanied by other company duties in the GDPR—including establishing data protection officers, using data protection impact assessments, and following the principles of data protection by design—this regime, if enforced, has the potential to be a sea change in how algorithmic decision-making is regulated in the EU.").

22's requirements.[45] Much more persuasively, others argue that human involvement must be meaningful, as the A29WP guidance states, involving a person who has the "authority and competence to change the decision."[46] Article 22 in fact creates a strong presumption, or even prohibition, against solely automated decision-making, subject to three exceptions.[47] The GDPR intends to target decision-making systems that are fully automated, those which are, for instance, wholly constituted by a machine learning model. The right to explanation applies to these cases only.[48]

Articles 13, 14, and 15 then require that the data controller provide information about "the logic involved" in the automated decision-making. Here again, the language of the text itself is ambiguous. It is likely that this involves a requirement to explain the logic of the whole machine learning model rather than a subset of the predictions it produces.[49] If so, the GDPR is broader than other legal requirements to explain automated decisions, such as the requirement in the Equal Credit Opportunity Act (ECOA) that an applicant be provided a "statement of specific reasons for the action taken."[50] The ECOA requirement focuses on the individual outcome only, while the GDPR arguably requires a broader form of explanation.

This would seem to produce a view of the resulting right to explanation that falls squarely within the 1A category. The GDPR, on this view, requires an explanation of the logic of an entire machine learning model,

---

[45] *See* Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, *supra* note 32, at 88 ("Quite crucially, this creates a loophole whereby even nominal involvement of a human in the decision-making process allows for an otherwise automated mechanism to avoid invoking elements of the right of access.").

[46] *See* Casey et al., *Rethinking Explainable Machines*, *supra* note 1, at 171 ("According to the A29WP, companies must ensure that any human 'oversight of [a] decision is meaningful, rather than just a token gesture' if they intend for their systems to fall outside the scope of Article 22's provisions pertaining to decisions 'based solely on automated processing.'").

[47] These exceptions are: consent, contract, or if authorised by Union or member state law. *See* Kaminski, *The Right to Be Explained, Explained*, *supra* note 1, at 197-198 (describing the three exceptions to the Article 22 right and prohibition); Isak Mendoza and Lee A. Bygrave, *The Right Not to Be Subject to Automated Decisions Based on Profiling* 14 (U. OSLO FAC. L. LEGAL STUDIES, Research Paper No. 2017-20, 2017) (providing the exceptions to the Article 22 right).

[48] "Interpreting Article 22 as a prohibition rather than a right to be invoked means that individuals are automatically protected from the potential effects this type of processing may have." Article 29 *Data Protection Working Party, Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation* 2016/679 (Feb. 6, 2018), at 20 [hereinafter A29WP]. Also note that, according to the Guidelines, the exceptions in Article 22 should be interpreted narrowly. *Id.* at 13.

[49] *See* Selbst & Powles, *Meaningful Information and the Right to Explanation*, *supra* note 23, at 236.

[50] 12 C.F.R. § 1002.9(a)(2)(i).

where that model constitutes the whole decision-making procedure that results in legal or similarly significant effects on a data subject.

This is not only a limited reading of the provisions and intent of the GDPR, it also completely misunderstands the role that explanation should play in a broader system for structuring accountability in the governance of algorithmic decision-making. Machine learning is a way of establishing a decision-making procedure. It is best thought of as a process, one that involves choices at every stage. These choices are made by institutions who design and integrate machine learning models into their decision-making procedures. These choices profoundly shape the form the machine learning model takes, the role it plays in their decision-making procedures, and the effects those decisions have on individuals over time. We believe that the RtE should be read in the context of the GDPR's broader provisions for mechanisms to secure accountability over time. These focus on the ex-ante design and implementation of decision-making procedures using machine learning.[51]

There are three crucial choices in the process of machine learning itself that must be considered, along with a broader set of choices about the role the machine learning model plays in the decision-making procedure, and the policy or commercial aims the institution has in deploying it.

1. Outcome of Interest

First, the outcome of interest is what the machine learning model looks for, that is, what it predicts, ranks, or classifies. The selection of an outcome of interest very often embeds important moral and political choices, which profoundly shape the predictions, classifications, or rankings the model will produce.[52] This choice, and the reasons for making it, require justification.

2. Training Data

Second, the training data set is what the machine learning models from. Recent research has developed several technical approaches to the evaluation of fairness in training data.[53] There are multiple aspects to the selection and construction of a training dataset, all of which can be

---

[51] *See infra* note 52.

[52] *See generally* Cary Coglianese and David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L. J. 1147 (2017).

[53] *See generally* Rich Zemel et al., *Learning Fair Representations*, 2013 INT'L CONF. MACH. LEARNING 325 (2013); J. Henry Hinnefeld et al., *Evaluating Fairness Metrics in the Presence of Dataset Bias*, CORNELL UNIV. (Sept. 24, 2018), http://arxiv.org/abs/1809.09245.

extremely important in shaping the predictions of the resulting machine learning model. These range from choices about time periods, demographic representativeness, and how to label the data.

3. Features

Third, the features included in a machine learning model. This includes choices about whether to include or exclude protected traits, such as race and gender. Removing a protected trait from a model is neither necessary nor sufficient to prevent discrimination in machine learning. In fact, preventing discrimination may require that information about individual membership of protected groups be *included* in machine learning models; fairness might require awareness, not blindness.[54] It also includes choices about whether to simplify the model by reducing the number of variables.[55]

Accountability requires justification. The form of explanation that justification requires depends on who is justifying what to whom. The GDPR is concerned with holding to account institutions which use automated decision-making procedures in important spheres.[56] Technical explanations of the logic of a machine learning model to an isolated individual will not be conducive to the kind of ongoing accountability the GDPR requires. The very form a machine learning model takes depends on choices made by humans in its design and implementation. The notion of providing a technical explanation of a machine learning model completely obscures the important and prior question: How did the rules that govern the operation of the automated decision come to be what they are? That is a question about the justification of institutional choices which is both prior to and much more significant than the question of what the

---

[54] *See generally* Cynthia Dwork et al., *Fairness through Awareness*, 2012 PROC. INNOVATIONS. IN THEORETICAL COMPUT. SCI. 214 (2012); Talia B. Gillis and Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459, 471 (2019); Symposium, Nina Grgic - Hlaca et al., *The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making*, 29 CONF. ON NEURAL INFO. PROCESSING SYS. (2016).

[55] Veale et al. describe a case in which the performance target of 75 percent was specified in advance, so the number of features could be reduced from 18,000 to 200, then 20, then 8, "because it's important to see how it works, we believe." Michael Veale, Max Van Kleek, and Reuben Binns, *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making* 440, PROC. OF THE 2018 CHI. CONF. ON HUM. FACTORS IN COMP. SYS. (2018), http://arxiv.org/abs/1802.01029.

[56] For instance, the 'spheres' in which DPIAs might be required are described as 'high-risk' in the text. The A29WP guidance lists a set of concrete criteria that make clear the broad scope of what 'high-risk' might mean. *See* Casey et al., *supra* note 1, at 176 ("Article 35(7) of the GDPR enumerates four basic features that all DPIAs must, at a minimum, contain."); A29WP, *supra* note 48.

rules are. It is also, we have argued, a question to which the GDPR's provisions aim to elicit an answer.

This is precisely what the A29WP guidance states. The guidance explains that "the complexity of machine-learning" algorithms "can make it challenging to understand how an automated decision-making process or profiling works."[57] Such complexity, the guidance continues, "is no excuse for failing to provide information."[58] Companies whose decisions are subject to the provisions of Article 22 "should find simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision," "not necessarily a complex explanation of the algorithms used or [a] disclosure of the full algorithm."[59] The guidance further clarifies that this will include information used in the decision-making process, including: categories of data; the source of that information; how many profiles were constructed and used in the procedure; and how that profile is used for a decision about the data subject.[60]

Institutions always make choices about how to design and integrate machine learning models into their decision-making procedures. In these choices lie trade-offs about discrimination and fairness, who wins and who loses, along with a host of other normative and epistemological assumptions. It is for these choices that an institution must be held accountable. The GDPR's provisions for a RtE must be understood in this context. Surrounding guidance makes clear that the appropriate form of explanation is not specifically about the logic of the machine learning model, but the choices an institution made in designing and integrating it into their decision-making system.[61]

B. *To Whom Should the Justification Be Offered: Regulators and Citizens*

There is also confusion about to whom the justification is owed. Here again, the language of the GDPR is not helpful. The GDPR text itself does not explain the aims of a RtE. However, the guidelines explain that "the data subject will only be able to challenge a decision or express their view

---

[57] *Id.* at 25.

[58] *Id*.

[59] *Id*.

[60] *Id*. at 31.

[61] For a useful overview of the kinds of choices that might be required for the form of justification at which the GDPR aims, which they term 'legibility,' *See generally* Gianclaudio Malgieri and Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, *supra* note 6.

if they fully understand how it has been made and on what basis."[62] The emphasis on the ability to challenge the decision reflects the fact that on this view, the purpose of the explanation is to invoke a data subject's other fundamental rights. As Kaminski puts it, "[i]ndividual transparency provisions, as the guidelines make clear, are intended to empower individuals to invoke their rights under the GDPR."[63]

We think this is a problem not with scholarly interpretations of the GDPR, but with the reasoning of the text itself and the guidelines surrounding it. The idea that the disclosure of information produces the enforcement of rights is not supported by evidence. Other areas of consumer behaviour research suggest people often struggle understanding straightforward information about products and how they pertain to their personal information.[64] The GDPR's instrumental individual transparency approach goes one step further, assuming that individuals will not only understand the information they are provided, but also that they will recognize violations of their legal rights and act on them.[65] Furthermore, many of the fundamental concerns about using machine learning to make decisions – most notably those related to bias and discrimination – can only be understood with a systematic and aggregate analysis of the decision-making procedure. The explanation of an individual decision to an isolated individual will not enable this kind of aggregate analysis; in fact, it may even obscure demands for obtaining it. The GDPR's account of the instrumental aim of an individual RtE is not convincing.

If systemic accountability is placed front and centre, rather than individual rights, it is clear that institutional justification of decision-making procedures must be offered to empowered, well-resourced regulators. There are ample provisions in the GDPR for doing just this. The individual RtE should not distract or detract from these provisions for systemic accountability. Rather, as we have consistently argued, the RtE should be viewed as a means to this broader end.

A focus on systemic accountability produces a very different view of the kind of explanations a regulator might require from an institution. We believe that at minimum, an explanation that supports the form of justification required by systemic accountability would answer the following questions. In all cases, the institution must not only provide a satisfactory answer to the question, it must provide reasons for the answers

---

[62] A29WP, *supra* note 48, at 27.

[63] Kaminski, *The Right to Be Explained, Explained*, *supra* note 1, at 211.

[64] *See supra* note 24. *See also* Oren Bar-Gill and Kevin Davis, *(Mis)perceptions of Law in Consumer Markets*, AM. L. & ECON. REV. (2017) (discussing misperceptions of the law, which is an additional reason that disclosures alone may be insufficient).

[65] *See* Edwards & Veale, *Enslaving the Algorithm*, *supra* note 5, at 52.

given. Where relevant, answers could be accompanied by quantitative data and analysis.

1. *What are the goals of the decision-making procedure?*

2. *What are the company policies that constrain or inform the decision-making procedure, including the role machine learning plays within it?*

3. *How did the company define the outcome of interest the machine learning model was trained to predict? Why?*

4. *How did the company select and construct the data on which the model was trained? If relevant, how was the data labelled and by whom? Was the impact of using other training data considered?*

5. *What features did the company choose to include or excluded in the model? Why?*

6. *Does the decision-making procedure involve human discretion? How precisely do the automated and human element of the decision-making procedure interact? Has the company considered how this interaction effects aggregate outcomes?*

7. *Has the lender considered how this interaction affects decisions?*[66]

The GDPR has ample mechanisms for encouraging, if not requiring, companies to answer these questions. As Kaminski argues, rather than "arguing over" the "instrumental value of individual notice, or publicly releasing source code," we should be debating how to obtain structured "accountability across a firm's decision-making, over time."[67]

Consider Data Protection Impact Assessments (DPIAs).[68] DPIA's are a "process for building and demonstrating" compliance by systematically

[66] For an alternative and insightful list of questions, s*ee generally* Malgieri and Comandé, *supra* note 6, at 29-30.

[67] Kaminski, *Binary Governance, supra* note 1, at 35.

[68] There are others mechanisms in the GDPR for attaining systemic accountability, such as auditing and ethical review boards. *See e.g.* Kaminski, *Binary Governance, supra* note 1, at 8 ("The instrumental rationale for regulating algorithmic decision-making counsels that regulation should try to correct these problems, often by using systemic accountability mechanisms, such as ex ante technical requirements, audits, or oversight boards, to do so."); Kroll et al., *supra* note

examining how automated decision-making procedures are designed and implemented. They are meant to be an "iterative process" that fall within the GDPR's broader "data protection by design" principles, which apply throughout the design, implementation and monitoring of a decision-making procedure.[69] DPIAs are more than simple recommendations of best practice. They are intended to apply to a broad range of institutions which use data to make important decisions. Importantly, those decisions must not be solely automated. As the A29WP guidance states, DPIAs apply "in the case of decision-making including profiling with legal or similarly significant effects that is not wholly automated, as well as solely automated decision-making defined in Article 22(1)."[70] Where appropriate, companies should "seek the views of data subjects or their representatives" during the DPIA process.[71] And companies should explain their reasons for making the choices they did in the design and implementation of their models.

In this context, the scope and content of the RtE is much broader. As Casey et al. argue, the right to explanation "is no mere remedial mechanism to be invoked by data subjects on an individual basis, but implies a more general form of oversight with broad implications for the design, prototyping, field testing, and deployment of data processing systems."[72] We agree with Veale and Edwards that *ex ante* DPIAs will "become the required norm for algorithmic systems, especially where sensitive personal data, such as race or political opinion is processed on a large scale."[73]

This is as it should be. The form of explanation required for institutional justification will often not be the technical explanation of the logic of machine learning models to isolated individuals. This is the 1A

33, at 660 ("Beyond transparency, auditing is another strategy for verifying how a computer system works."); Selbst & Barocas, *supra* note 3, at 1133 ("The most common trigger of the latter is a lawsuit, in which documents can be obtained and scrutinized and witnesses can be deposed or examined on the stand, but auditing requirements are another possibility.").

[69] *See* Casey et al. *supra* note 1, at 172-173; A29WP, *supra* note 48, at 29 ("As a key accountability tool, a DPIA enables the controller to assess the risks involved in automated decision-making, including profiling. It is a way of showing that suitable measures have been put in place to address those risks and demonstrate compliance with the GDPR.").

[70] *Id*. at 32 ("The following list, though not exhaustive, provides some good practice suggestions for controllers to consider when making solely automated decisions . . . ."). *See also* Casey et al., *supra* note 1, at 174 (According to the Regulation, DPIAs are mandatory '[w]here a type of processing[,] taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons.'") (internal citations omitted).

[71] *Id*. at 36.

[72] *Id.* at 39.

[73] Edwards & Veale, *Slave to the Algorithm?, supra* note 5, at 78.

category. Rather, it should be an explanation of the decisions an institution made in the design of a machine learning model and its integration into their decision-making procedure, to an empowered regulator. This is the 2B category. Reporting to a regulator rather than to an individual is necessary to reveal aggregate patterns and effects that are not discoverable when considering a decision in isolation.[74] Regulators and other public bodies have the technical knowledge, skills and time to evaluate information that an individual does not.[75] The very purpose of regulators is to take actions in situations when it is individually not worthwhile, but is socially desirable.

CONCLUSION

The RtE debate should begin with the foundational goal: accountability. Accountability is constitutive of democratic self-governance. It is an integral aspect of a citizenry's ongoing authorization of the complex decision-making systems which shape their lives. Part of what it means to be a citizen of a self-governing polity is to give and receive justifications of those decision-making systems. Explanations are

[74] One of us has written about this type of aggregate analysis elsewhere when considering the type of information a lender would provide to the CFPB to allow testing of whether credit pricing algorithms are compliant with discrimination law. *See generally* Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459 (2019). In the context of credit pricing discrimination, this has been one of the most significant barriers to a successful discrimination complaint. The passing of the Home Mortgage Disclosure Act in 1975, increased the ability to bring a successful discrimination claim and class action against lender since the Act mandated the disclosure of mortgage applications and their outcomes, allowing for an aggregate consideration of mortgage decisions. *See e.g.* Robert G. Schwemm & Calvin Bradford, *Proving Disparate Impact in Fair Housing Cases after Inclusive Communities*, 19 N.Y.U. J. LEGIS. & PUB. POL'Y 685, 713-15 (2016).

[75] Future technical research into explainability and interpretability in machine learning could benefit from assuming that the appropriate audience for their approaches is not isolated individuals but regulators. The great strength of Dwork's 'individual fairness' approach is that it isolates the normative choices and therefore makes possible a form of accountability, e.g. fair affirmative action, through the choice of the distance metric. It can require access to protected status information during the design phase, usually explicitly prohibited, which may require a big shift in policy. What matters though is a *procedure* which justifies the choice of the distance metric, which can be explained to either a regulator or, in some cases, those who are actually subject to the decision. *See* Dwork et al., *supra* note 54, at 2 (describing the "[c]onnection between individual fairness and group fairness," Dwork et al. state that "[s]tatistical parity is the property that the demographics of those receiving positive (or negative) classifications are identical to the demographics of the population as a whole. Statistical parity speaks to group fairness rather than individual fairness, and appears desirable, as it equalizes outcomes across protected and non-protected groups."); *see also id*. at 3 ("Justifying the *availability* of or access to the distance metric in various settings is one of the most challenging aspects of our framework, and in reality the metric used will most likely only be society's current best approximation to the truth.").

valuable insofar as they are required to achieve systemic accountability over time. In practice, this means that the appropriate form of explanation will depend on who is justifying what to whom. We have argued that the RtE debate focuses far too much on the explanation of the logic of a machine learning model to isolated individuals. What matters for accountability is the justification by an institution of the choices it made in the design and implementation of a machine learning model. The form of systemic accountability should drive the form of institutional justification, which in turn, should drive the appropriate form of explanation.

Interpreting the GDPR matters because it is likely to shape future regulation of algorithmic decision-making. The primary concerns that arise when using machine learning to make, or assist with, important decisions are not satisfactorily addressed by focusing on the rights of isolated individuals, or the logic of an individual machine learning model itself. As we develop comprehensive governance structures to address the concerns that arise from the use of machine learning in decision-making, we should move beyond frameworks that rely on the individual enforcement of rights, and towards those which develop a systemic approach to establishing and maintaining accountability within a complex modern democracy.

This means moving beyond privacy as a lens through which to view the governance of algorithmic decision-making. Some of the limited ways in which the GDPR has been interpreted have been transplanted from older debates about privacy. This is partly because the GDPR itself grew out of earlier privacy provisions and it is partly because scholars who interpret it often cut their cloth in the privacy field. The focus on individual rights, as well as the notice and consent framework that underpins the GDPR's approach, are all characteristic of approaches to addressing concerns about privacy. As Kaminski puts it, "the strong system of individual rights" within the GDPR may come "at the cost of correcting systemic problems essential for achieving accountability in modern democracies."[76] If the RtE is interpreted as requiring explanations of the logic of machine learning models to isolated individuals, these explanations are not likely to be useful to regulators in evaluating whether to accept the justification of an institution of its decision-making procedure. That is, such

[76] Kaminski, *Binary Governance*, *supra* note 1, at 74. This also means relating current discussions about the governance of algorithmic decision-making to a rich literature on regulatory strategies in an administrative state. *See e.g. id*. at 30-31 ("If there is already concern in administrative law over insulating government bureaucrats from electoral and judicial oversight, collaborative governance compounds such concerns by involving private parties.")

explanations may actually obstruct systemic accountability. Most challenging of all, the GDPR requires companies to assist in the enforcement of citizens' fundamental rights. This effectively privatizes the protection of individual rights. The GDPR and the literature surrounding it has no satisfactory account of how its provisions are to be subject to democratic oversight. Accountability matters because it is constitutive of collective self-government. Future regulatory provisions must focus more directly on developing mechanisms within modern democracies that can secure accountability in the governance of algorithmic decision-making systems.

We are currently in a moment of choice. We are choosing how to integrate humanity's most powerful decision-making tool – machine learning – into a range of complex human activities. We have argued that institutional justification, not algorithmic explanation, is essential to the accountability constitutive of democratic self-government. The technical explanation of machine learning models is never sufficient, is often not necessary, and sometimes actively distracts from, the justification of the decision-making systems of which they are a part. We must think through what it means to reason about the justifications an institution should offer for its choices in how and why it constructed its decision-making procedure in the way it did – that is, a justification of why the rules are what they are. We have offered a sketch of what such a system of reasoning might look like.

We must keep our eyes on the right prize. That prize is accountability. Institutional power is held in check by other institutions with the authority and resources sufficient to hold them to account. To attain that prize requires a laser-like focus on choice in the face of apparent technical inevitability. In this case, it means requiring institutions to justify their choices about how they have constructed their decision-making systems. Not being distracted by whizzy technical explanations of their machine learning models work – or even, of that most dangerous of terms, artificial intelligence.